Supporting Information

Lin et al. 10.1073/pnas.1614654114

SI Materials and Methods

Shotgun Metagenomic Sequencing and Data Analysis. To obtain sufficient DNA for shotgun metagenomic sequencing, multiple displacement amplification was performed using the GenomiPhi V2 DNA Amplification Kit (GE Healthcare) following the manufacturer's instructions. Briefly, 1 μL of DNA was used as the template and was mixed with 9 μL of sample buffer. The mixed DNA was heated at 95 °C for 3 min and cooled to 4 °C, before incubation at 30 °C for 90 min with 1 μL of enzyme mixture and 9 μL of reaction buffer. To terminate the reaction, the sample was heated at 65 °C for 10 min. For each sample, nine amplifications were pooled to reduce potential bias. These were purified using TIANquik Maxi Purification Kit (Tiangen).

Shotgun sequencing of metagenomic DNA was performed using Illumina HiSeq 2000 using the pair-end 125 × 125 library with a 600-bp inset size (Beijing Genomics Institute, Beijing, China). The entire dataset of two samples is ∼5.55 Gb. Illumina reads were trimmed to remove the adapter sequences and low-quality bases, after which 86–88% of paired reads were retained for each sample. Trimmed, paired-end reads were assembled using a multiple *k*-mer–based assemblies (64). Briefly, metagenomic reads of each sample were individually assembled into contigs using the Velvet, version 1.2.10, assembler (49) with a range of *k*-mers (41, 51, 61, 71, 81, and 91). The different assembles were subsequently merged, and the duplicated and suboptimal contigs were removed through CD-HIT-EST (65) using a sequence identity threshold of 0.95 and a word length of 8 to get the final assembly for each sample. Resulting contigs were filtered by a minimal length cutoff of 1 kb.

Population Genome Binning of a Magnetotactic *Nitrospirae* from HCH. Contigs of sample HCH were sorted using BLASTn alignment against the NCBI genomes database (version May 2015) together with previously sequenced MTB draft genomes of Mcas (17), Mchi (18), and Mbav (18). BLASTn alignment hits with *E* values larger than 1×10^{-5} were filtered, and the taxonomical level of each contig was determined by the lowest common ancestor algorithm implemented in MEGAN, version 5 (50). All contigs binned to known *Nitrospirae* MTB species of Mcas, Mbav, and Mchi were selected. Due to the incomplete nature of available magnetotactic *Nitrospirae* draft genomes, the remaining contigs were further classified using CLARK, version 1.1.2 (51), based on reduced sets of *k*-mers by comparison with available genomes or draft genomes of MTB strains. The measure of conservation of gene content and gene order of MGCs between HCH-1 and three available *Nitrospirae* MTB (Mcas, Mbav, and Mchi) is the ratio between the number of genes located in conserved content and order and the total number of bidirectional best-hits genes between MGCs of HCH-1 and three *Nitrospirae* MTB.

Implicit Phylogenomic Analysis of *Nitrospirae* MTB Genes. The global implicit phylogenetic pattern of the magnetotactic *Nitrospirae* genomes of HCH-1, Mcas, Mbav, and Mchi was inferred using HGTector 0.2.0 (53). Protein sequence similarity search was performed using DIAMOND 0.9.7 (66) against a database (generated by HGTector) that contains one representative per species from all available nonredundant RefSeq prokaryotic proteomes (October 2015), plus the MTB proteomes reconstructed in this study. Quality cutoffs for valid hits were *E* value ≤ 1e-20, percentage identity ≥ 30%, and query coverage ≥ 50%. For each protein-coding gene, the top 250 highest-scoring hits from different species were retained. For each hit, a "relative bit score" was calculated as the original bit score of the hit divided by the bit score of the query sequence aligned against itself. The overall distribution pattern of all genes in a genome was visualized by plotting the sum of the bit scores of hits within phylum *Nitrospirae* against that outside this phylum per gene.

Divergence Time Estimation. Molecular-dating analyses were performed using PhyloBayes, version 4.1c (63). The CAT-GTR model was implemented for amino acid replacement, and analyses were run under either the log-normal autocorrelated relaxed clock (-ln) or the uncorrelated gamma multipliers (-ugam). For each condition, two replicate chains with 20,000 generations were run. Dates were assessed by running the readdiv with the first 20% of generations removed as burn-in for each analysis. Two different combinations of age constraints were used for the divergence time estimation. For the first combination of age constraints, the minimum age of the root of Oxyphotobacteria (oxygenic Cyanobacteria) was set at 2.32 Ga (the rise in atmospheric oxygen) (67), and the maximum age was set at 3.0 Ga (40, 68). For the second combination, a minimum age of 1.9 Ga (the first widely accepted fossil oxygenic Cyanobacteria) (69) and a maximum age of 2.32 Ga (postdating the rise of oxygen according to ref. 70) were implemented as the oxyphotobacterial root. In addition, for the second combination another time constraint, the divergence time between Oxyphotobacteria and Melainabacteria, was included, which was set from 2.5 Ga (70) to 3.8 Ga (the end of late heavy bombardment). For all analyses, the age calibration for the last common ancestor of all taxa used in this study was set between 2.32 and 3.8 Ga (71).

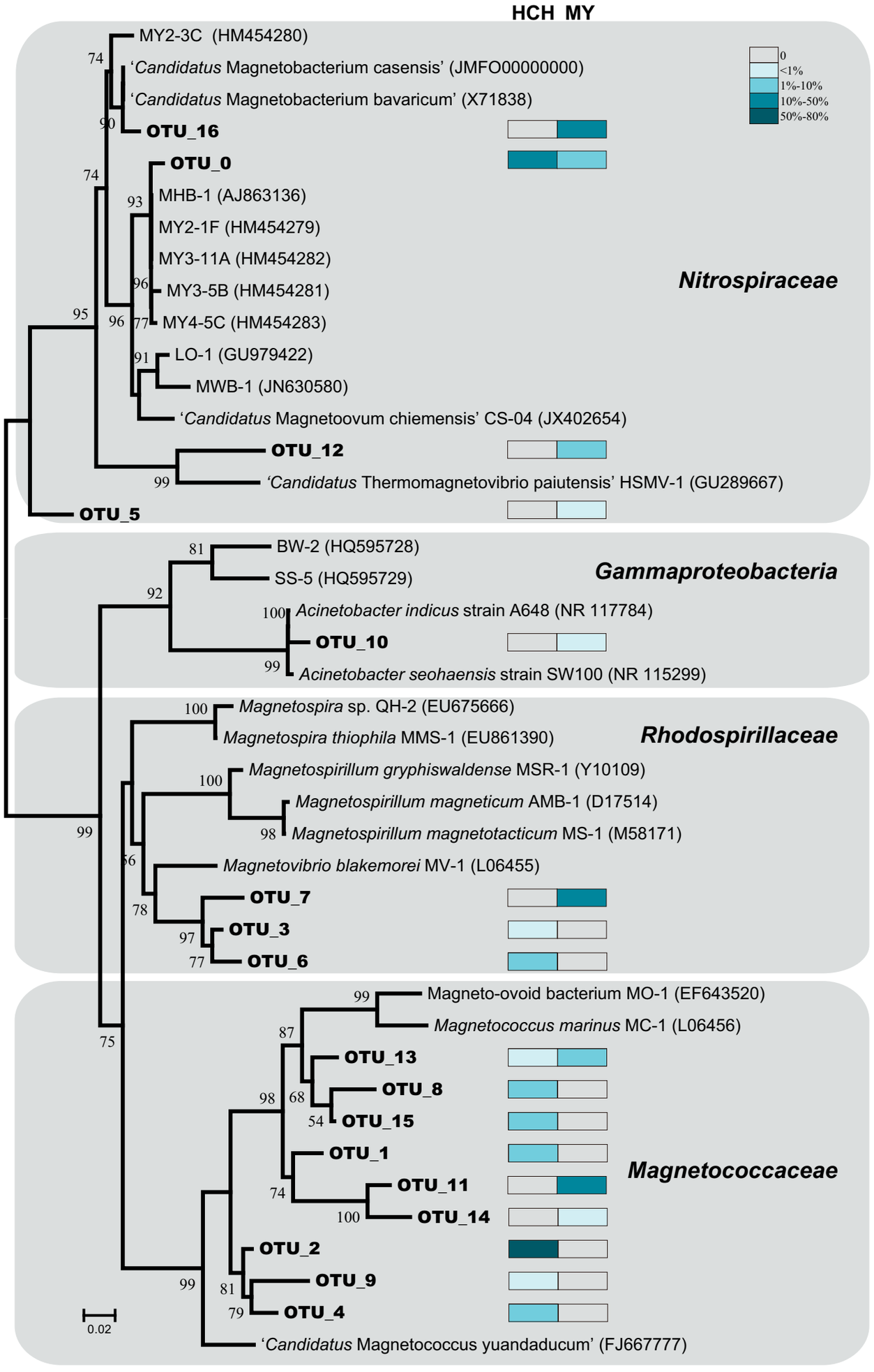

Fig. S1. Phylogenetic tree of operational taxonomic units (OTUs at 97% threshold similarity) for 16S rRNA gene clone libraries of MTB communities from the city moat of Xi'an in Shaanxi province (HCH) and Lake Miyun near Beijing (MY). The evolutionary history was inferred by using the maximum-likelihood method based on the Kimura two-parameter model with 100 bootstraps. On the right-hand side, a heatmap shows the relative abundance and distribution of each OTU from this study.

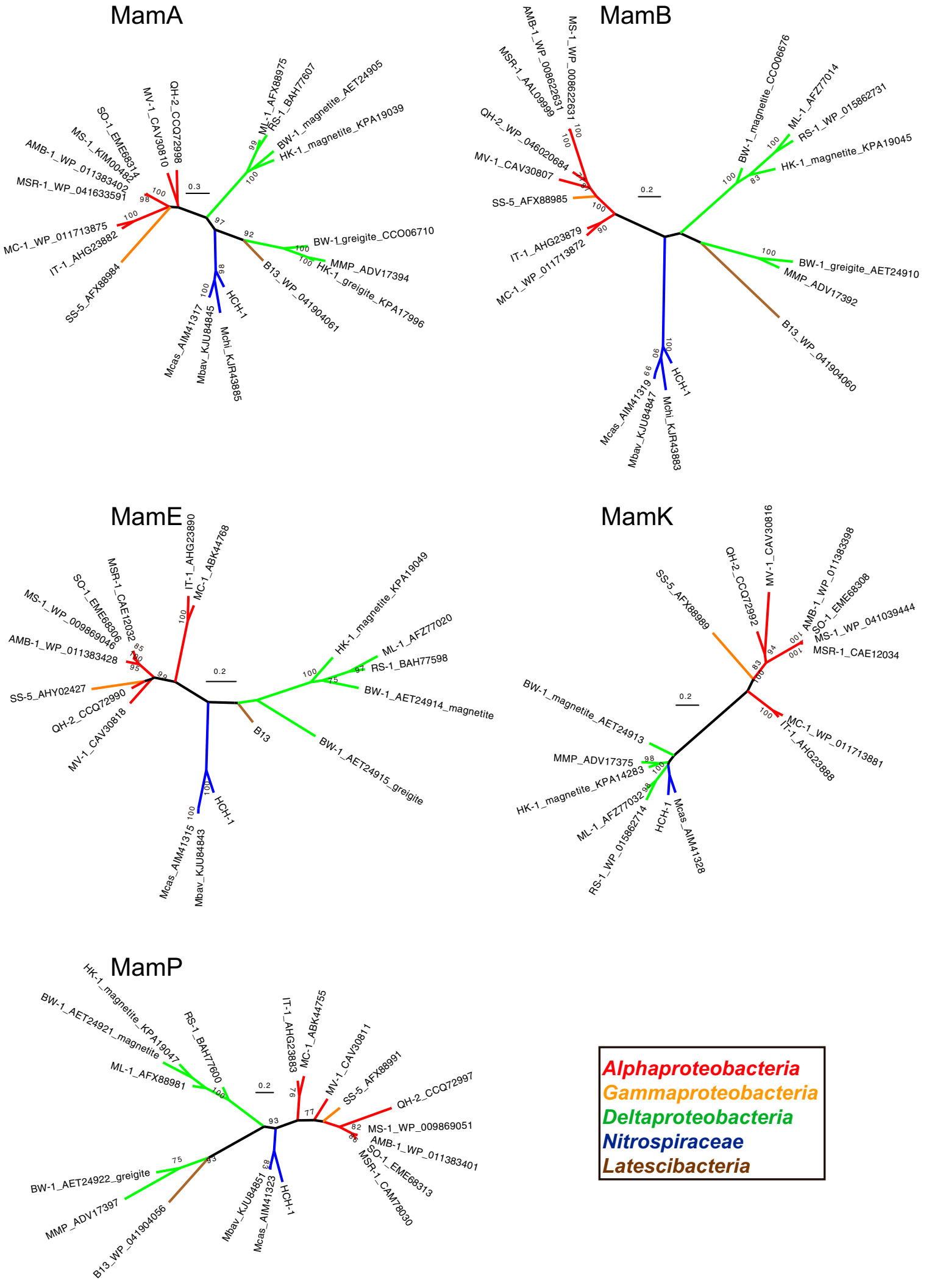

Fig. S2. Bootstrap consensus trees of five magnetosome proteins (MamABEKP) based on the maximum-likelihood method. Only full-length protein sequences were included in this analysis. Bootstrap values are expressed as percentages, and only values of more than 75% are shown. MSR-1, *Magnetospirillum gryphiswaldense* MSR-1; AMB-1, *Magnetospirillum magneticum* AMB-1; SO-1, *Magnetospirillum* sp. SO-1; QH-2, *Magnetospira* sp. QH-2; MC-1, *Magnetococcus marinus* MC-1; BW-1, *Candidatus* Desulfamplus magnetomortis BW-1; MMP, *Ca*. Magnetoglobus multicellularis; RS-1, *Desulfovibrio magneticus* RS-1; ML-1, alkaliphilic magnetotactic strain ML-1; MV-1, *Magnetovibrio blakemorei* MV-1; IT-1, *Ca.* Magnetofaba australis strain IT-1; SS-5, Gammaproteobacteria magnetotactic strain SS-5; HK-1, *Ca*. Magnetomorum sp. HK-1; B13, *Latescibacteria bacterium* SCGC AAA252-B13.

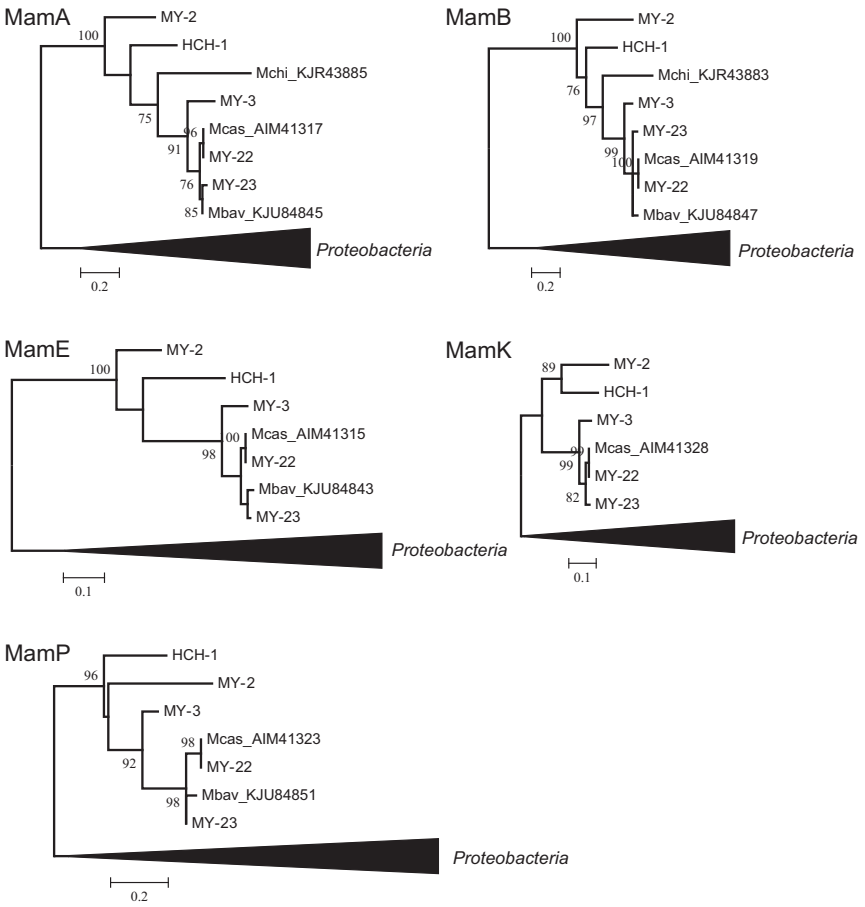

Fig. S3. Bootstrap consensus trees of magnetosome proteins MamABEKP from the four additional *Nitrospirae* MGCs of MY and those full-length proteins from all available *Nitrospirae* MTB based on the maximum-likelihood method. Bootstrap values are expressed as percentages, and only values of more than 75% are shown.

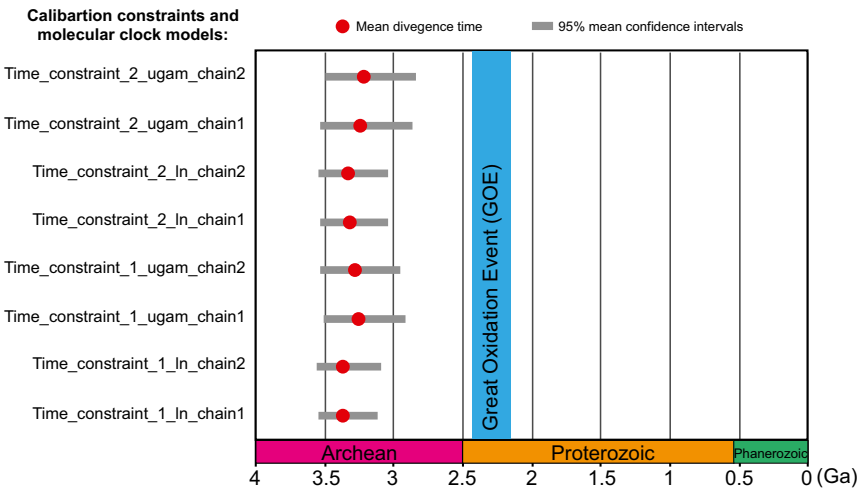

Fig. S4. Summary of mean divergence dates for the *Nitrospirae* and *Proteobacteria* phyla estimated using Bayesian relaxed molecular-clock analyses with two different time constraints and two different molecular clock models (see *SI Materials and Methods* for details). Two replicated chains were run for each condition. The input phylogenomic tree used here is shown in Fig. S5.

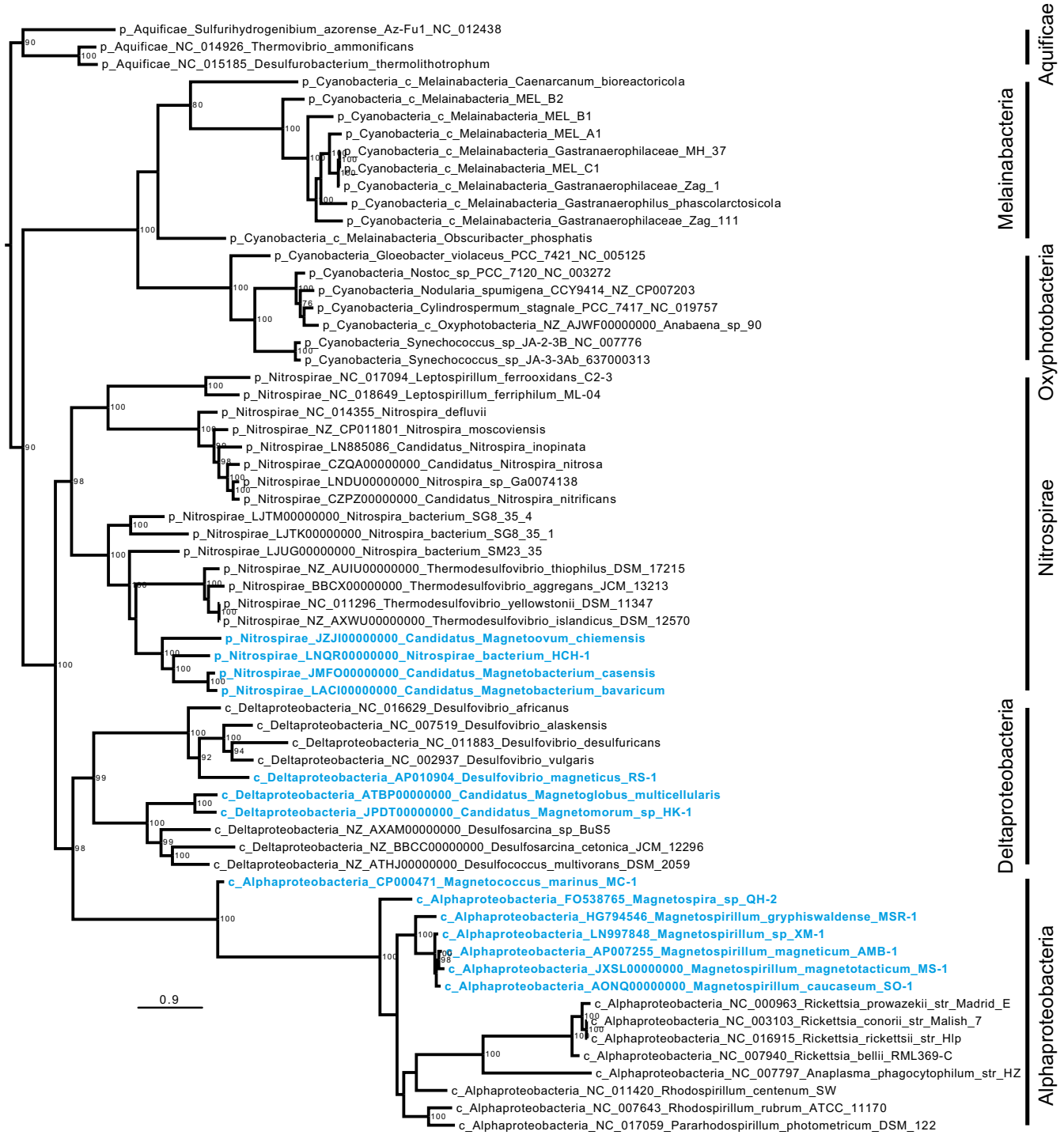

Fig. S5. Phylogenomic maximum-likelihood tree of 64 bacterial genomes. Bootstrap values are expressed as percentages, and only values of >75% are shown. Magnetotactic bacteria are displayed in blue. The Aquificae strains were used as outgroup.

Other Supporting Information Files

Table S1 (DOCX)
Table S2 (DOCX)
Table S3 (DOCX)